10. Li, H.B. *et al*. (2008) *Halomonas korlensis* sp. nov., a moderately halophilic, denitrifying bacterium isolated from saline and alkaline soil. *Int. J. Syst. Evol. Microbiol.* 58, 2582–2588.
11. Sánchez-Porro, C. *et al*. (2009). Description of *Kushneria aurantia* gen. nov., sp. nov., a novel member of the family *Halomonadaceae*, and a proposal for reclassification of *Halomonas marisflavi* as *Kushneria marisflavi* comb. nov., of *Halomonas indalinina* as *Kushneria indalinina* comb. nov. and of *Halomonas avicenniae* as *Kushneria avicenniae* comb. nov. *Int. J. Syst. Evol. Microbiol.* 59, 397–405.
12. Kuhnert, P. *et al*. (2010) *Basfia succiniciproducens* gen. nov., sp. nov., a new member of the family *Pasteurellaceae* isolated from bovine rumen. *Int. J. Syst. Evol. Microbiol.* 60, 44–50.
13. Gregersen, R.H. *et al.* (2009) Comparative studies on [*Pasteurella*] *testudinis* and [*P.*] *testudinis*-like bacteria and proposal of *Chelonobacter oris* gen. nov., sp. nov. as a new member of the family *Pasteurellaceae*. *Int. J. Syst. Evol. Microbiol.* 59, 1583–1588.
14. Spergser, J. *et al.* (2010) *Mycoplasma mucosicanis* sp. nov., isolated from the mucosa of dogs. *Int. J. Syst. Evol. Microbiol.* (in press) DOI:10.1099/ijs.0.015750-0.
15. ICSB Subcommittee on the Taxonomy of Mycoplasmatales (1972) Proposal for minimal standards for descriptions of new species of the order Mycoplasmatales. *Int. J. Syst. Bacteriol*. 22, 184–188.
16. Logan, N.A. (2005) International Committee on Systematics of Prokaryotes; Subcommittee on the taxonomy of the genus *Bacillus* and related organisms: Minutes of the meeting, 30 July 2002, Paris, France. *Int. J. Syst. Evol. Microbiol.* 55, 977–979.
17. Tindall, B.J. *et al*. (2006) Valid publication of names of prokaryotes according to the rules of nomenclature: past history and current practice. *Int. J. Syst. Evol. Microbiol.* 56, 2715–2720.
18. Holmes, B. and Farmer III, J.J. (2009) International Committee on Systematics of Prokaryotes Subcommittee on the taxonomy of *Aeromonadaceae*, *Vibrionaceae* and related organisms. *Int. J. Syst. Evol. Microbiol.* 59, 2638–2640.

Biography

Martha E Trujillo is a senior lecturer in the Department of Microbiology and Genetics at the University of Salamanca (Spain) and Vice-Dean of the Faculty of Agricultural and Environmental Sciences. Martha undertook a PhD in Bacterial Systematics at the University of Newcastle upon Tyne, UK, followed by a postdoctoral position in the Department of Microbiology at Novartis AG in Switzerland. She joined the University of Salamanca in 2002, where she currently carries out her research on plant–bacteria interactions with special emphasis on the relationship between *Micromonospora* and nitrogen-fixing nodules of legumes. In addition, she is interested in bacterial communities that thrive in deteriorated rock monuments.

Under the **Microscope**

Prokaryotic taxonomy in the sequencing era and the role of MLSA in classification

Peter Kämpfer

Institut für Angewandte Mikrobiologie
Justus-Liebig-Universität Giessen
Heinrich-Buff-Ring 26, D-35392 Giessen, Germany
Email peter.kaempfer@umwelt.uni-giessen.de

Stefanie P Glaeser

Institut für Mikro- und Molekularbiologie
Justus-Liebig-Universität Giessen
Heinrich-Buff-Ring 26, D-35392 Giessen, Germany
Email peter.kaempfer@umwelt.uni-giessen.de

The initial step in prokaryote species and genera descriptions is now largely based on the 16S rRNA gene sequencing approach followed often by a very restricted additional phenotypic characterisation of the representatives of the potential novel taxa. Despite the advantages of the sequence-based approaches, there appears to be a tendency to classify new species on the basis of comparative sequence analyses of 16S rRNA gene sequences and other gene sequence data (multilocus sequence analyses, MLSA), contrary to the indications of other data. However, the biological meaning behind these sequence data is not always clear, and one should be careful with comprehensive taxonomic rearrangements until there is better insight of these data.

Organisms can be classified in different ways, but the resulting classification systems remain abstract ideas (mental representations). In biology, the ultimate goal of taxonomy is to establish a system that mirrors the "order in nature". The term "natural" is now most often associated with evolution. In prokaryote microbiology, the taxonomic concepts try to mirror the origin of life, hence the cell as the basic unit of life remains the level of consideration.

The analysis of the small subunit ribosomal RNA (16S rRNA) gene has clearly revolutionised prokaryotic taxonomic studies. For the first time, a hierarchical taxonomy on the basis of a practical molecular marker was possible.

The advantages and the pitfalls of 16S rRNA sequence-based

studies have been already summarised by Ludwig and Klenk[1]. One important step, the alignment, can be critical at variable regions if deletions or insertions have occurred and sometimes it is very difficult to indicate the homologous bases. For the 16S rRNA (gene) sequences (and for sequences of protein coding genes), the functional pressure dictates the preservation of the sequence responsible for the three-dimensional structures of the final gene products and hence the probability of base substitutions leading to functional products. The limitations of tree reconstructions, the different information content, problems of plesiomorphy, functional constraints, multiple 16S rRNA gene copies with small intragenomic differences (up to 2 or even 5%) and so on have also been thoroughly addressed by Ludwig and Klenk[1] and others, but despite these problems, 16S rRNA gene sequence-based comparisons have been and are still invaluable in describing the breadth of prokaryotic diversity and are indispensable to allocate an unknown to a taxon.

One major critical point is the resolution power of the 16S rRNA gene sequence below the genus level, because of its conserved structure. Organisms sharing very similar or even identical 16S rRNA sequences may be more diverse at the whole genome level, than those having more variable positions[2], and it is also very important to note that the branching patterns at the periphery of trees (sometimes at the genus level, but most often at the species level) cannot reliably reflect phylogeny in the sense of common ancestry (independent from the "treeing algorithm"). Here it should be noted briefly, that 16S rRNA sequence comparisons are often simply taken as to be "phylogenetic"[3]. Sneath[4] has, however, provided a clear argument, that these analyses were clearly phenetic and not cladistic. Hence many so-called "phylogenetic" trees, published in the taxonomic literature, are often based on simple similarity calculations and are phenetic and not cladistic.

An example is shown in Figure 1 for the genus *Acinetobacter*. Here the 16S rRNA gene sequence in general is clearly not reliable for the identities at the species level. This is the case for the majority of bacterial and archaeal genera, especially those harbouring many species.

But what is in this context a species, or better, what reflects best the mental representation "species"?

It is agreed overall that the taxonomic category "species" represent the fundamental units in taxonomy. Despite the different approaches to define this unit, any species definition should be pragmatic, operational and universally applicable, and should serve the whole community[8]. Stackebrandt *et al.*[9] provided in 2002 a species circumscription as "a category that circumscribes a (preferably) genomically coherent group of individual isolates/strains sharing a high degree of similarity in (many) independent features, comparatively tested under highly standardised conditions".

At this time large-scale genomic sequencing was beyond the imagination of most biologists and hence the description of a species was based on the 70% DNA-DNA hybridisation (DDH) standard already introduced by Wayne *et al.*[10], which was for a long time pragmatic and universally applicable within the bacterial domain of life[4,8]. This method has often been criticised for being difficult to implement due to the cumbersome DDH experiments and the high experimental error. The advantages and pitfalls have been reviewed in detail by Rosselló-Mora[11].

With the development of new sequencing methods, it has now become feasible to generate gene and genome sequences in a relatively short period of time. One very important development in the analysis of bacterial genomes for taxonomic purposes was the introduction of the average nucleotide index[12-16].

Current available comparisons of DDH results of genome sequences show 95% average nucleotide identity of shared genes (ANI), which correlates well with the 70% DDH value defining bacterial species[13-16]. As outlined by Cole *et al.*[12] multilocus sequence analysis (MLSA) represents an "intermediate resolution" of the 16S rRNA gene and genome-based approach to resolve the "phylogenetic" resolution at the species level. MLSA (see, for example,[17]) is based on multilocus sequence typing (MLST) an approach to microbial typing based upon the sequences of multiple genes, (often housekeeping genes) that was first introduced by Maiden *et al.*[18]. Housekeeping genes are preferably used for MLST analysis because they are expected to evolve at a slow and constant rate.

For the majority of bacterial pathogens, MLST websites are now available (http://www.mlst.net/): the most comprehensive databases exist for *Campylobacter jejuni* and *Neisseria meningitidis*. MLST analysis in epidemiological studies is not based on the calculation of phylogenetic relationships but is rather a cluster analysis based on "allelic profiles", whereas similarities between different alleles are not considered. "Phylogenetic" calculations using the sequence information behind the MLST schemes (referred as MLSA) can be used for bacterial identification and classification, because it offers the opportunity to incorporate the insights available from population genetics and phylogenetic approaches into bacterial systematics.

The application of MLSA as a method to replace DNA-DNA hybridisation is also associated with problems, as shown here briefly for the genus *Acinetobacter*. *Acinetobacter* comprises strictly aerobic, non-motile, Gram-negative, oxidase-negative and catalase-positive diplococcoid rods. The delineation of species within the genus *Acinetobacter* has resulted in a complex situation. Extensive investigations on strains belonging to the genus *Acinetobacter* using the DNA–DNA pairing studies have resulted in the recognition of several DNA–DNA homology groups (genomic species). The genus comprises now 22 species with validated names and 11 genomic species that have not been named until now. The majority of them are difficult to differentiate phenotypically. On the basis of 16S rRNA gene sequencing studies they show similarities between 93.5% and

99.1% (Figure 1) but without a clear species allocation. The branching patterns shown in the tree cannot reliably reflect phylogeny in the sense of common ancestry.

The genus *Acinetobacter* is also homogeneous with respect to phenotypic traits. For these reasons, molecular identification methods are increasingly developed and being validated against DNA–DNA hybridisation data. These methods include also the study of "housekeeping" genes, for example, those encoding RNA polymerase subunit B (*rpoB*), gyrase subunit B (*gyrB*), or the recA protein (*recA*). Especially for the clinically relevant species of the *Acinetobacter baumanni* group (comprising *A. calcoaceticus*, *A. baumannii*, *Acinetobacter* genomic species 3 and *Acinetobacter* genomic species 13) even more comprehensive MLST schemes have been developed, including fragments of the *gltA, gyrB, gdhB, recA, cpn60, gpi* and *rpoD* genes (for example,[19]). These MLST data clearly show even a high diversity among the *A. baumannii* isolates and hence the high resolution of MLST, even below the species level.

A comprehensive application of MLSA analysis on the whole genus *Acinetobacter* has not been applied so far. A detailed comparison of the *rpoB* sequences of the type (and other) strains of all *Acinetobacter* species (Figure 2) shows similarities of 77.3 to 99.0%. The derived amino acid sequence data (Figure 2) reveal much higher similarities of 84.6 to 100%. The same situation was found for the *gyrB* sequence similarities (ranging from 75.1 to 100%) in comparison to the derived amino acid sequence similarities (84.8 to 100%, data not shown) as well as for the *recA* sequence similarities (80 to 97.4%) in comparison to the amino acid sequence similarities (96.2 to 100%, data not shown).

Branching patterns obtained with the *rpoB* sequences clearly show the differences to the 16S rRNA gene-based tree and the branching patterns based on the *gyrB* and the *recA* sequences again showed differences. Furthermore, amino acid sequence-based trees partially show different branching patterns than corresponding gene sequence-based trees (for example, for *recA*).

Of course, one possibility to overcome the problem of contrasting branching patterns is the combination of sequences from protein coding gene sequences with 16S rRNA gene sequences or sequences of genes that shared the same resolution as the 16S rRNA-based analysis, which was done, for example, by Serrano *et al.*[24] who applied MLSA to resolve taxonomic conflicts in genus *Marichromatium* (anoxygenic phototrophic bacteria of the family *Chromatiaceae*). But this may be highly problematic, because the resulting differences may not reflect the result of very different genetic (evolutionary) processes.

Concatenation is often recommended to study the sequences of more genes, but this may lead to additional problems[25]. In a simulation by Kubatko and Degnan[25], the concatenation approach produced a high level of discord among individual gene trees and led to a statistically inconsistent estimation. In addition, they

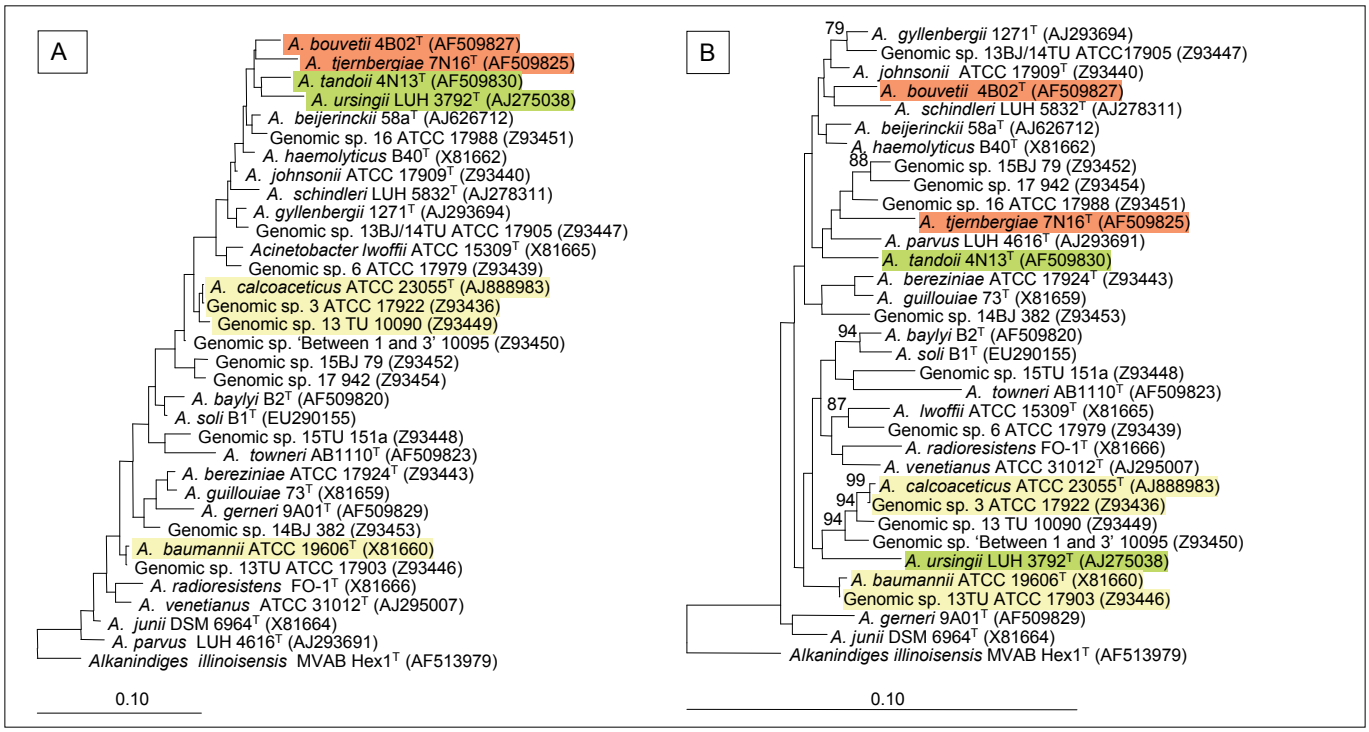

Figure 1. Phylogenetic analysis based on 16S rRNA gene sequences available from the European Molecular Biology Laboratory data library (accession numbers in parentheses). Trees were constructed using the ARB software package (version December 2007[5]) and the corresponding SILVA SSURef 100 database (release August 2009[6]).
A. Tree building was performed using the maximum likelihood method with fastDNAml[7] without conservatory filter.
B. Tree building was performed with the neighbour-joining method without conservatory filter. Bar, 0.10 nucleotide substitutions per nucleotide position.
Some branches are highlighted to point to selected differences.

found that the use of the bootstraps to measure support for the inferred phylogeny may lead to moderate to strong support for an incorrect tree under these conditions.

These selected examples clearly show that sequence data have to be interpreted very carefully as similarity at the phenotypic level (here the amino acid sequence is regarded as the phenotype) is very much higher than the underlying genotypic data.

With the investigation of genome data, however, new questions and problems are becoming obvious, which have not been sufficiently resolved until now. Given the enormous fluidity of the genomes, the exact rate and magnitude of gene exchange, which occurs over a broad range of taxonomic levels, it is expected, that a general definition of the category "species" may not be adequate given the expected and already documented genomic differences and the derived evolutionary consequences.

All genes in the genome, including the 16S rRNA genes, are subject to LGT (lateral or horizontal gene transfer) and recombination events and these events may differ significantly within different bacterial groups[26]. The majority of bacteria contain in their genomes specific genetic elements with encoding functions responsible for their own LGT and they may integrate genes or operons into the genome by site-specific recombination. MLST is an important method to study LGT (see[26] for more details). As a consequence, these investigations are important for MLSA analyses. As a recommendation, genes that have been largely affected by LGT should perhaps not be considered in comparative MLSA studies for taxonomic purposes.

This statement may be extended also to the analysis of whole genome sequences and, despite the advantages in this area, many questions remain unanswered as summarised by Konstantinidis *et al.*[14,15] and Cole *et al.*[12]. In this regard, an important fact is that genes and a genome do not exist on their own. In taxonomy (and also in many other biological disciplines) the living unit, that is, the cell should be the correct level of consideration.

Although molecular analysis can provide an enormous amount of data, we are far from able to interpret these data. There are numerous questions, for example, "Which genes belong to the conserved genome core are considered probably useful to define a taxon and which belong to accessory dispensable genetic elements?" The "overall" impact of processes such as LGT, gene duplication, recombination and rearrangements of genes in the genome is not clear and may vary considerably in different lineages (for example,[27-29]). In addition, the presence

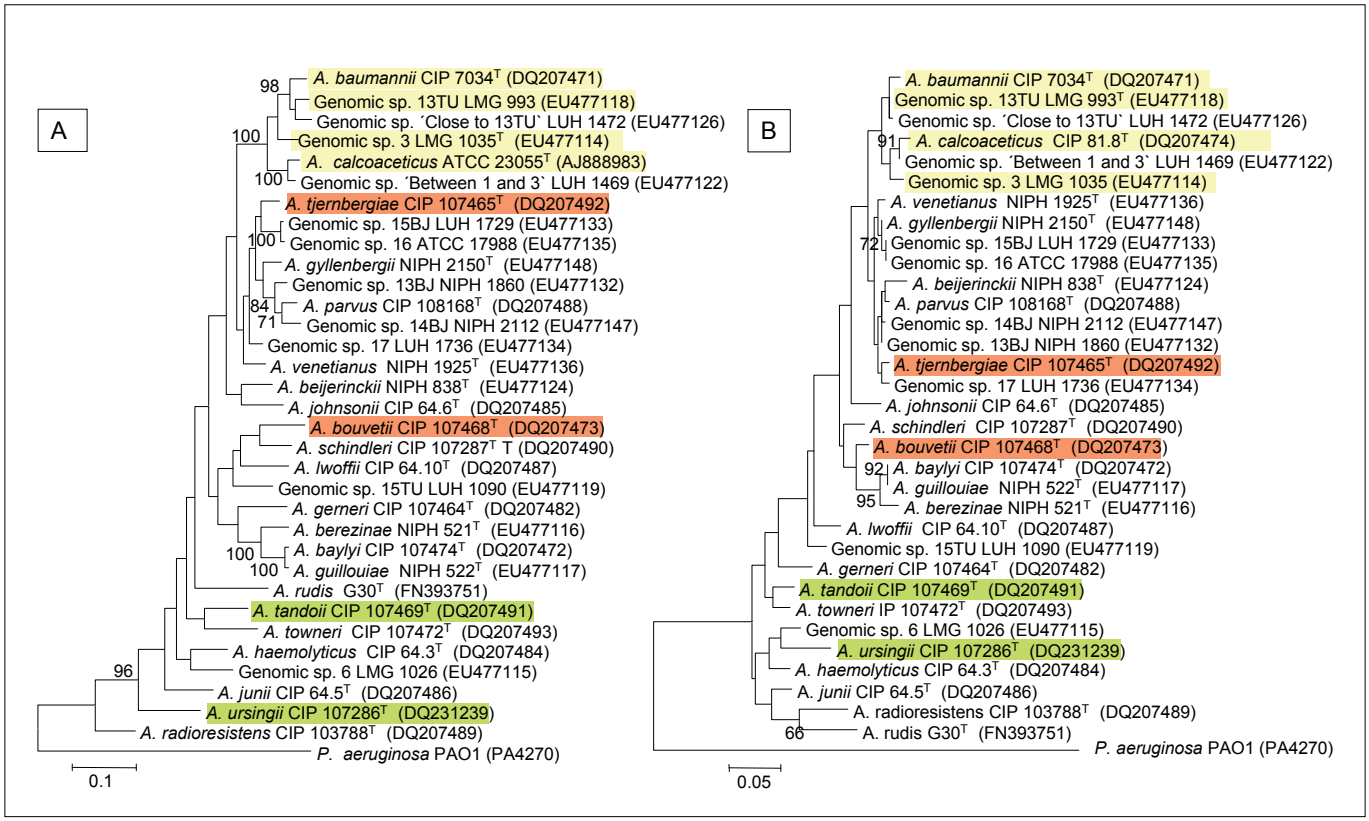

Figure 2. Phylogenetic resolution of *Acinetobacter* species and genomic species based on partial *rpoB* sequences (A) and corresponding amino acid sequences (B). Phylogenetic trees were generated with the Neighbour-Joining method[20]. Evolutionary distances were computed using the Kimura 2-parameter method (for gene sequences[21]) and the JTT matrix-based method (for amino acid sequences) and are in the units of the number of base/amino acid substitutions per site. All codon positions were included in the gene sequence analysis. All positions containing gaps and missing data were eliminated. A total number of 842 nucleotide and 280 amino acid positions were in the final data set. Numbers at branching points represent the percentage of replicate trees that confirm respective branching points in the bootstrap test (1000 replications)[22]. Only bootstrap values ≥70 are shown. The genome sequences strain *Pseudomonas aeruginosa* PAO1 was used as the out-group. Analyses were conducted in MEGA5[23]. Some branches are highlighted to point to selected differences.

of genes and gene clusters (whether expressed or "silent") can have a totally different biological meaning, and the roles of structural elements (some of them phenotypically recognisable by the so-called "chemotaxonomic" methods) and biochemical pathways (also recognised by studying the phenotype at different levels) should be consistent with the underlying genetic data, which is essentially the aim of a "polyphasic taxonomy" as originally defined by Colwell *et al*.[30]. Despite the advantages in sequencing, the task of describing novel taxa is one that still requires careful selection and use of a wide variety of methodologies. The availability of an increasing number of sequenced genomes from a diverse range of prokaryotes is providing a wealth of new data, which have to be interpreted with care. Nevertheless, experience shows that only the interplay between genetic and phenotypic data sets provides a sound basis for appreciating and describing the diversity of prokaryotes and has the potential to become the foundation of a more stable, in-depth taxonomy of the prokaryotes[31].

References

1. Ludwig, W. and Klenk, H.-P. (2001) Overview: a phylogenetic backbone and taxonomic framework of prokaryotes. In *Bergey´s Manual of Systematic Bacteriology* (2nd edn) (Garrity, G.M. ed.), pp. 49–65, Springer, New York.
2. Stackebrandt, E. *et al.* (1994) Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int. J. Syst. Bacteriol*. 44, 846–849.
3. Woese, C.R. (1987) Bacterial evolution. *Microbiol. Rev*. 51, 221-271.
4. Sneath, P.H.A. (1989) Analysis and interpretation of sequence data for bacterial systematics – the view of a numerical taxonomist. *Syst. Appl. Microbiol*. 12, 15–31.
5. Ludwig, W. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res.* 32, 1363–1371.
6. Pruesse, E. *et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* 35, 7188–7196.
7. Olsen, G.J. *et al.* (1994) fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *CABIOS.* 10, 41–48.
8. Rosselló-Mora, R. and Amann, R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev.* 25, 39–67.
9. Stackebrandt, E. *et al.* (2002) Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 52, 1043–1047.
10. Wayne, L.G. *et al.* (1987) International Committee on Systematic Bacteriology. Report of the *ad hoc* committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* 37, 463-464.
11. Rosselló-Mora, R. (2006) DNA-DNA reassociation methods applied to microbial taxonomy an their critical evaluation. In *Molecular Identification, Systematics, and Population Structure of Prokaryotes* (Stackebrandt ed.), pp. 23–50, Springer-Verlag, Heidelberg, Berlin, Germany.
12. Cole, J.R. *et al.* (2010) Microbial Diversity and Phylogeny: Extending from rRNAs to Genomes In *Environmental Molecular Microbiology* (Liu, W.-T. and Jackson, J.K., eds), pp. 1–19, Caister Academic Press.
13. Goris, J. *et al.* (2007) DNA-DNA hybridization values and their relation to whole genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91.
14. Konstantinidis, K.T. and Tiedje, J.M. (2005) Genomic insights into the species definition for prokaryotes. *Proc. Natl. Acad. Sci. USA.* 102, 2567–2572
15. Konstantinidis, K.T. *et al.* (2006a) The bacterial species definition in the genomic era. *Phil. Trans. R. Soc. Lond. B. Biol. Sci.* 361, 1929–1940.
16. Konstantinidis, K.T. *et al.* (2006b) Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl. Environ. Microbiol.* 72, 7286–7293.
17. Gevers, D. *et al.* (2005) Opinion: Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* 9, 733–739.
18. Maiden, M.C. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic organisms. *Proc. Natl. Acad. Sci. USA.* 95, 3140–3145.
19. Wisplinghoff, H. *et al.* (2008) Molecular epidemiology of clinical *Acinetobacter baumannii* and *Acinetobacter* genomic species 13TU isolates using a multilocus sequencing typing scheme. *Clin. Microbiol. Infect*. 14, 708–715.
20. Saitou, N. and Nei, M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
21. Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.
22. Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution.* 39, 783–791.
23. Tamura, K. *et al.* (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony *Methods Mol. Biol. Evol.* (submitted).
24. Serrano, W. *et al.* (2010) Evaluation of the use of multilocus sequence analysis (MLSA) to resolve taxonomic conflicts within the genus *Marichromatium*. *Syst. Appl. Microbiol.* 33, 116–121.
25. Kubatko, S.L. and Degnan, J.H. (2007) Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol*. 56, 17–24.
26. Giffard, P. (2010) Bioinformatics of microbial sequences. In *Infectious diseases informatics* (Sintchenko, V. ed), pp. 27–52, Springer-Verlag, New York.
27. Bapteste, E. *et al.* (2008) Lateral gene transfer challenges principles of microbial systematics. *Trends Microbiol*. 16, 200–207.
28. Bapteste, E. *et al.* (2009) Prokaryotic evolution and the tree of life are two different things. *Biology Direct* 4, 34 DOI: 10.1186/1745-6150-4-34
29. Dagan, T. *et al.* (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc. Nat. Acad. Sci. USA.* 105, 10039–10044.
30. Colwell, R.R. (1970) Polyphasic taxonomy of bacteria. In *Culture Collections of Microorganisms* (Iizuka, H. and Hazegawa, T., eds), pp. 421–436, University of Tokyo Press.
31. Tindall, B.J. *et al.* (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol*. 60, 249–266.

Biography

Professor Peter Kämpfer, microbiologist, is the current Director of the Institute of Applied Microbiology at the Justus-Liebig University Giessen. He received his PhD from the University of Bonn in 1991 and worked since 1995 in Giessen then on several bacterial groups, mainly from environmental samples. He was involved in the description of several novel bacterial taxa and is at present the Editor-in-Chief of the *International Journal of Systematic and Evolutionary Microbiology* and the Vice-Chairman of the Board of Trustees of Bergey´s Manual Trust. Current research activities comprise research on bacterial communities in indoor environments and the phyllosphere.

Stefanie P Glaeser studied nutritional sciences and did her Bachelor and Master thesis in the field of microbial ecology at the Institute of Applied Microbiology at the Justus-Liebig University, Giessen. Ever since she started to work with microorganisms she has been interested in their ecological role, isolation and taxonomic affiliation. She is currently preparing her PhD thesis at the Institute for Microbiology and Molecular Biology at the Justus-Liebig University in the field of freshwater bacterioplankton stress responses. Her current research interest looks at the isolation and description of ecologically significant microorganisms from freshwater environments and their use as laboratory model systems.